Towards Succinct and Relevant Image Descriptions

Desmond Elliott
Institute of Language, Communication, and Computation
School of Informatics
University of Edinburgh
d.elliott@ed.ac.uk

What does it mean to produce a *good* description of an image? Is a description good because it correctly identifies all of the objects in the image, because it describes the interesting attributes of the objects, or because it is short, yet informative? Grice's Cooperative Principle, stated as "Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975), alongside other ideas of pragmatics in communication, have proven useful in thinking about language generation (Hovy, 1987; McKeown et al., 1995). The Cooperative Principle provides one possible framework for thinking about the generation and evaluation of image descriptions.[1]

The immediate question is whether automatic image description is within the scope of the Cooperative Principle. Consider the task of searching for images using natural language, where the purpose of the exchange is for the user to quickly and accurately find images that match their information needs. In this scenario, the user formulates a complete sentence query to express their needs, e.g. *A sheepdog chasing sheep in a field*, and initiates an exchange with the system in the form of a sequence of *one-shot* conversations. In this exchange, both participants can describe images in natural language, and a successful outcome relies on each participant succinctly and correctly expressing their beliefs about the images. It follows from this that we can think of image description as facilitating communication between people and computers, and thus take advantage of the Principle's maxims of Quantity, Quality, Relevance, and Manner in guiding the development and evaluation of automatic image description models.

An overview of the image description literature from the perspective of Grice's maxims can be found in Table 1. The most apparent ommission is the lack of research devoted to generating minimally informative descriptions: the maxim of Quantity. Attending to this maxim will become increasingly important as the quality and coverage of object, attribute, and scene detectors increases. It would be undesirable to develop models that describe every detected object in an image because that would be likely to violate the maxim of Quantity (Spain and Perona, 2010). Similarly, if it is possible to associate an accurate attribute with each object in the image, it will be important to be sparing in the application of those attributes: is it relevant to describe "furry" sheep when there are no sheared sheep in an image?

How should image description models be evaluated with respect to the maxims of the Cooperative Principle? So far model evaulation has focused on automatic text-based measures, such as Unigram BLEU and human judgements of *semantic correctness* (see Hodosh et al. (2013) for discussion of framing image description as a ranking task, and Elliott and Keller (2014) for a correlation analysis of text-based measures against human judgements). The semantic correctness judgements task typically present a variant of "Rate the relevance of the description for this image", which only evaluates the description vis-à-vis the maxim of Relevance. One exception is the study of Mitchell et al. (2012), in which judgements about the ordering of noun phrases (the maxim of Manner) were also collected. The importance of being able to evaluate according to multiple maxims becomes clearer as computer vision becomes more accurate. It seems intuitive that a model that describes and relates every object in the image could be characterised as generating Relevant and Quality descriptions, but not necessarily descriptions of

[1] This discussion primarily applies to image *descriptions*, and not to image *captions*. See (Hodosh et al., 2013) and (Panofsky, 1939) for a discussion of the differences between descriptions and captions.

Proceedings of the 25th International Conference on Computational Linguistics, pages 109–111,
Dublin, Ireland, August 23-29 2014.

Category	Maxim	Attention in the literature
Quantity	Be as informative as required	???
	Do not be more informative than required	???
Quality	Do not say what you believe is false	All models exploit some kind of corpus data to construct descriptions that are maximally probable (Yang et al., 2011; Li et al., 2011; Kuznetsova et al., 2012; Le et al., 2013). These approaches typically use language modelling to construct hypotheses based on the available evidence, but may eventually be false.
	Do not say that for which you lack evidence	
Relevance	Be relevant	No models try to generate irrelevant descriptions. Dodge et al. (2012) explored the separation between what can be seen/not seen in an image/caption pair.
Manner	Avoid obscure expressions	No model has been deliberately obscure.
	Avoid ambiguity	Kulkarni et al. (2011) introduced visual attributes to describe and distinguish objects.
	Be brief	???
	Be orderly	Mitchell et al. (2012) and Elliott and Keller (2013) explicitly try to predict the best ordering of objects in the final description.

Table 1: An overview of Grice's maxims and the relevant image description models. ??? means that we are unaware of any models that implicitly or explicitly claim to address this type of maxim.

adequate Quantity. It is not clear that current human judgements capture this distinction, yet the gold-standard crowdsourced descriptions almost certainly do conform to the maxim of sufficient Quantity. A further important consideration is how to obtain human judgements for multiple maxims without making the studies prohibitively expensive.

Using Grice's maxims to think about image description from the perspective of enabling effective communication helps us reconsider the state of the art of automatic image description and directions for future research. In particular, we identified the open problems of determining the minimum and most relevant aspects of an image, and the challenges of conducting human evaluations along alternative dimensions to semantic correctness.

Acknowledgments

S. Frank, D. Frassinelli, and the anonymous reviewers provided valuable feedback on this paper. The research is funded by ERC Starting Grant SYNPROC No. 203427.

References

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alex Berg, and Tamara Berg. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772, Montréal, Canada.

Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, U.S.A.

Desmond Elliott and Frank Keller. 2014. Comparing Automatic Evaluation Measures for Image Description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 452–457, Baltimore, Maryland, U.S.A.

H. Paul Grice. 1975. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics 3: Speech Arts*, pages 41–58. Academic Press, Inc.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

E Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608, Colorado Springs, Colorado, U.S.A.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 359–368, Jeju Island, South Korea.

Dieu Thu Le, Jasper Uijlings, and Raffaella Bernardi. 2013. Exploiting language models for visual recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 769–779, Seattle, Washington, U.S.A.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, U.S.A.

K McKeown, J Robin, and K Kukich. 1995. Generating concise natural language summaries. *Information Processing & Management*, 31(5):703–733.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daum. 2012. Midge : Generating Image Descriptions From Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France.

Erwin Panofsky. 1939. *Studies in Iconology*. Oxford University Press.

Merrielle Spain and Pietro Perona. 2010. Measuring and Predicting Object Importance. *International Journal of Computer Vision*, 91(1):59–76.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK.